Тюнинг параметров TCP-соединений в высокоскоростных сетях

Красников
Валерий Викторович



План

- Планирование ресурсов оборудования
- Микросервисная архитектура и пропускная способность сети
- Покажи мне график нагрузки сети или что такое microburst
- Аппаратный буфер сетевого оборудования и причины перегрузок
- Что такое Incast и чем он грозит
- Решение проблемы Incast
- Немного теоретического отступления о реализации RTO TCP в Linux
- Настраиваем RTO. Проблема разных площадок
- Маленькие аккуратно разложенные сетевые грабли
- Что дальше?



Планирование ресурсов оборудования

- Планируем серверы и виртуальные машины
- Определяем потребности CPU/RAM/DISK
- Облако сети /10-40G хватит всем ☺
- Определяем потребности во внешних каналах
- Взлетаем……..

Не взлетели ☹

Оказывается, сети тоже надо планировать, не полагаясь на арендованную инфраструктуру или обещания вендора!



Микросервисы и сети

Как видим продукт мы

Internet

Frontend

MQ

Backend

DB



Микросервисы и сети

Как видит продукт сетевой инженер

Frontend
Frontend
Frontend
Frontend
Frontend
Internet
MQ
MQ
MQ
Backend
Backend
Backend
Backend
Backend
DB
DB
DB
switch



Покажи мне график нагрузки сети

Что такое microburst?

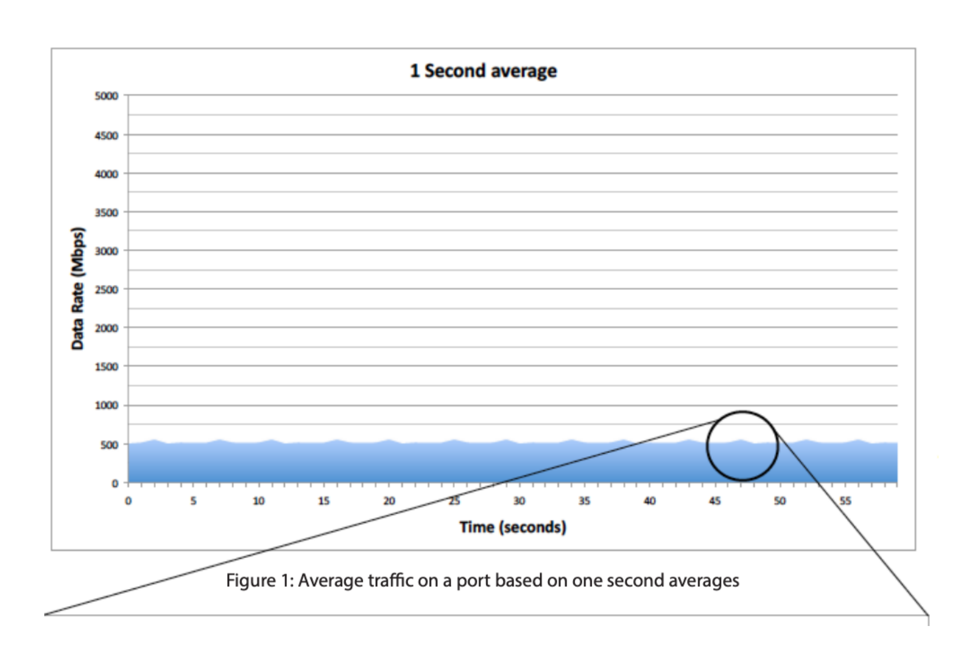


Figure 1: Average traffic on a port based on one second averages



Покажи мне график нагрузки сети

Что такое microburst?

1 Millisecond average

Data Rate (Mbps)

Time (milliseconds)

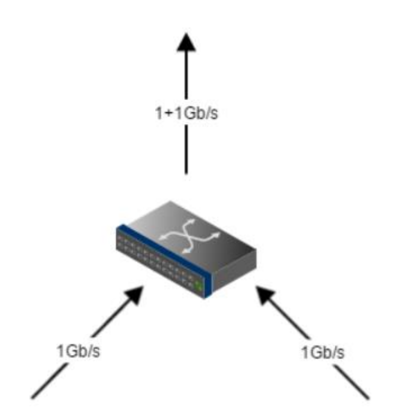
Figure 2: Microbursts – seen with finer time resolution

Microburst, длительностью 4 ms и пиком в 3.5 Gb



Покажи мне график нагрузки сети

Почему возникают microburst?





Аппаратный буфер сетевого оборудования

Для чего предназначен буфер:

- Сохранить тело пакета
- Сгладить поток пакетов до скорости выходного интерфейса
- Контролировать перегрузки (congestion)



Аппаратный буфер сетевого оборудования

Объемы буфера и tail drop:

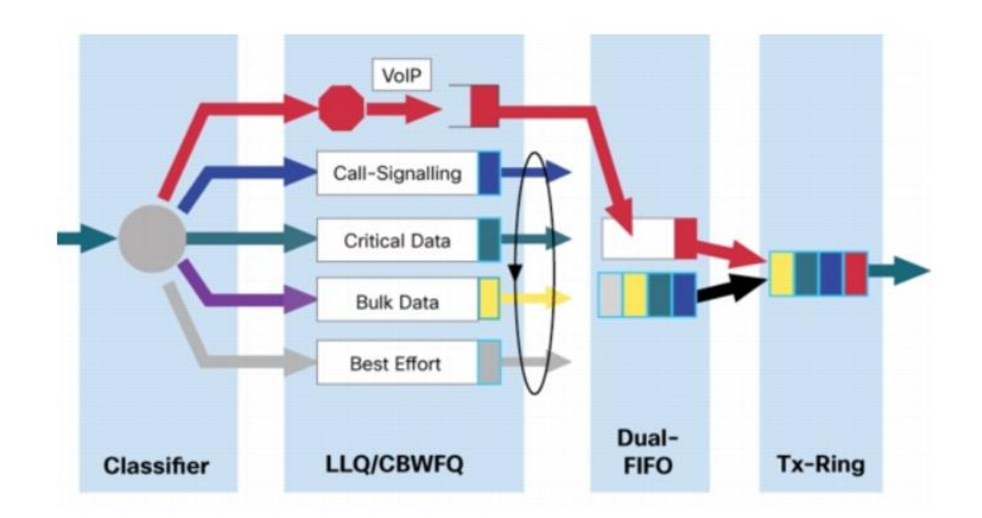
Time: 200 ms

- 1 Gb/s: (1000000000 * 0.2) / 8 = 25 MB
- 10 Gb/s: (10000000000 * 0.2) / 8 = 250 MB



Аппаратный буфер сетевого оборудования

Мы знаем, что делать – QoS!





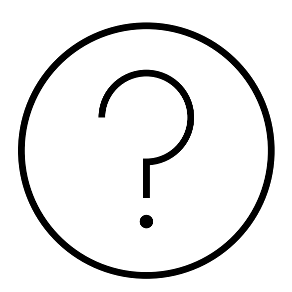
Причины перегрузок

- Трафик из высокоскоростного интерфейса должен попасть в менее скоростной (10 G->1 G)
- Трафик с нескольких входящих портов должен попасть в один исходящий DownLink- Backpressure
- Прочие типы burst-трафика
- **Incast**



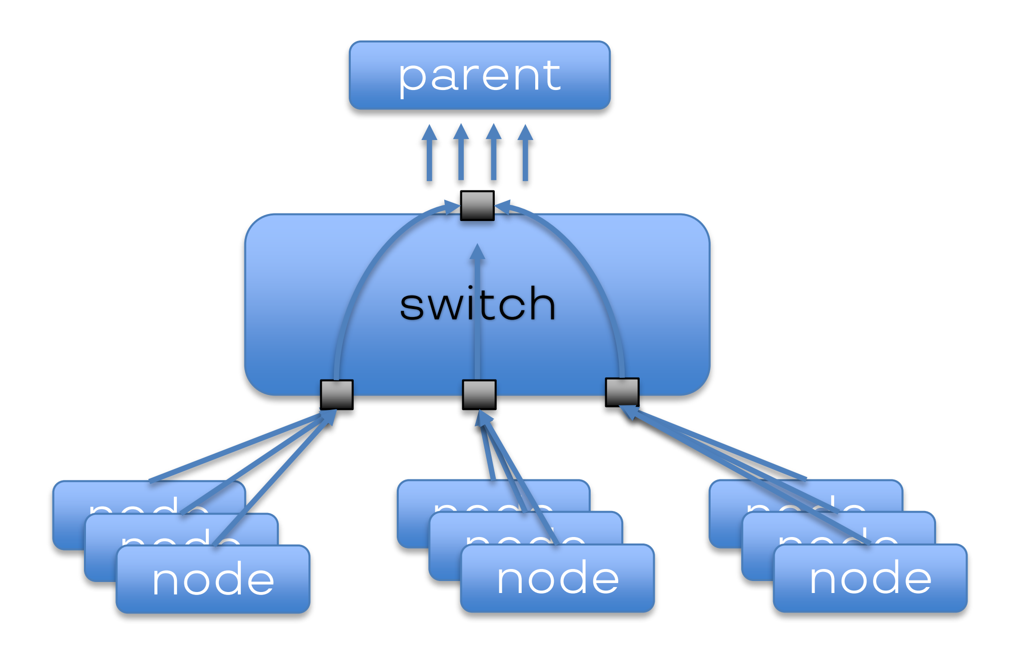
Причины перегрузок

Мы перегружены, попали под microburst,
QoS не вывозит, к нам пришел Incast





Что такое Incast и чем он грозит?



Hadoop, HDFS, Map Reduce, GFS

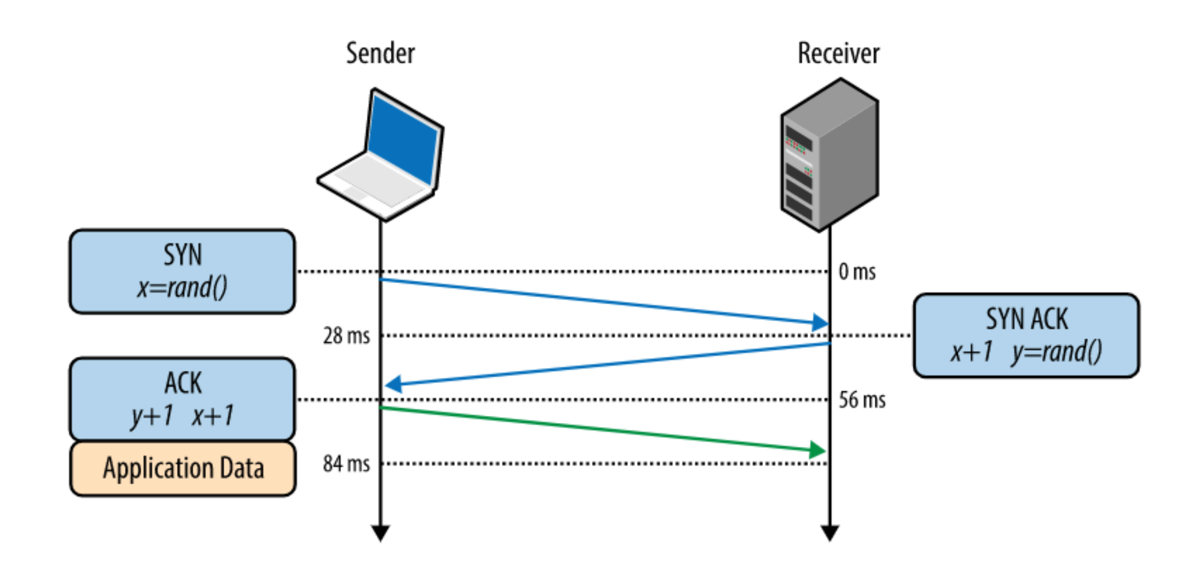


Определяем три решения проблемы Incast

- Большие буферы коммутатора
- Управление потоком
- Уменьшение минимального RTO TCP

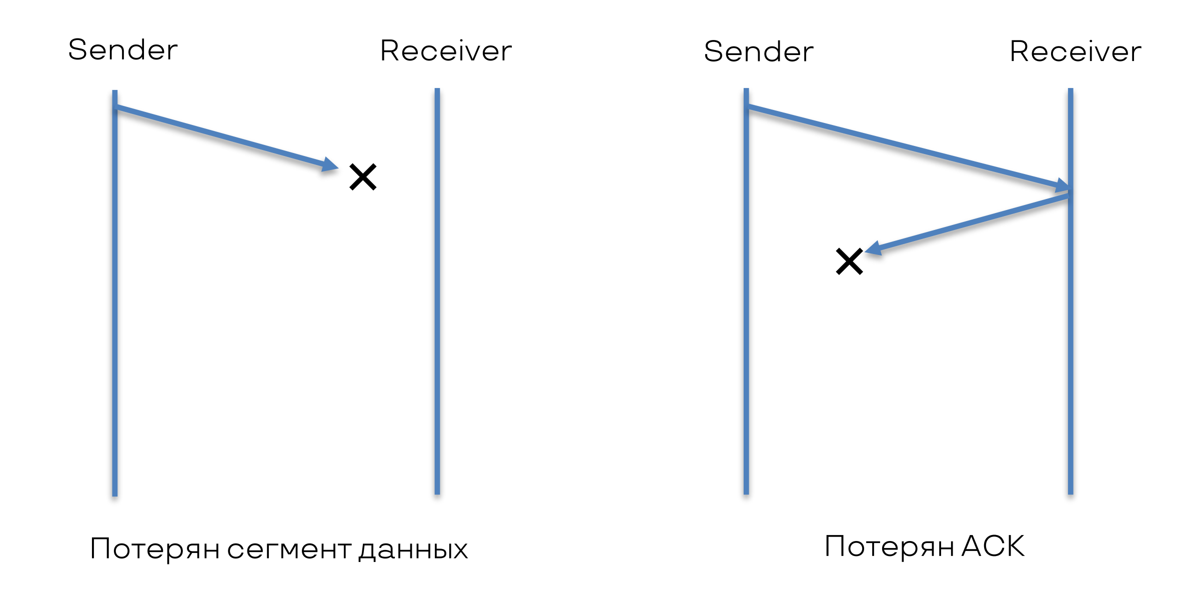


Немного теоретического отступления по протоколу TCP





Немного теоретического отступления по протоколу TCP





Немного теоретического отступления по протоколу TCP

Расчет значения RTO:

Sender Receiver

RTO

Retransmission

RTO мало
Ненужный повтор передачи

Sender Receiver

RTO

RTO большое
Медленная реакция на потери



Немного теоретического отступления по протоколу TCP

Расчет значения RTO:

RTTVAR <- (1 - beta) * RTTVAR + beta * |SRTT - R’|
SRTT <- (1 - alpha) * SRTT + alpha * R'

RTO <- SRTT + max (G, K*RTTVAR)

Источник: **RFC6298**



Немного теоретического отступления по протоколу TCP

В ядре Linux реализовано так же?

RFC - это рекомендации



Немного теоретического отступления по протоколу TCP

В ядре Linux реализовано так же?

[root@host test]# ss -i

cubic wscale:7,9 rto:202 rtt:0.567/0.271 ato:40 mss:1446 rcvmss:536 advmss:1448 cwnd:18 ssthresh:18 bytes_acked:21073 bytes_received:436 segs_out:23 segs_in:13 send 428.4Mbps lastsnd:15630243 lastrcv:15630246 lastack:15630242 pacing_rate 836.0Mbps rcv_space:28960
tcp ESTAB 0 0 ::ffff:10.11.0.1:19102 ::ffff:10.11.1.13:38484
cubic wscale:7,9 **rto:201 rtt:0.563/0.081** ato:40 mss:1448 rcvmss:536 advmss:1448 cwnd:12 ssthresh:9 bytes_acked:131502768 bytes_received:2507472 segs_out:99654 segs_in:44162 send 246.9Mbps lastsnd:5557 lastrcv:5593 lastack:5556 pacing_rate 493.0Mbps retrans:0/167 reordering:8 rcv_rtt:26208.2 rcv_space:29304



Пробуем решить проблему Incast своими силами

Меняем RTO(min)



Пробуем решить проблему Incast своими силами

Меняем RTO(min) или подкрутим f-rto?

```
# TCP stack tweaking for lossy wireless networks
net.ipv4.tcp_frto = 1
net.ipv4.tcp_frto_response = 2
net.ipv4.tcp_low_latency = 1
```




Пробуем решить проблему Incast своими силами

Меняем RTO(min)
Общая проблема всех выше преведенных решений

Мы **не можем** изменить глобальный минимальный RTO для TCP

Для разных типов оборудования в разных частях или разных ЦОД-ах может потребоваться разная настройка этого параметра



Пробуем решить проблему Incast своими силами

Меняем RTO(min)

```
[root@host test]# ip route

default via 10.111.3.254 dev ens192 proto static metric 100
10.0.0.0/8 via 10.111.3.254 dev ens192 proto bird metric 32

10.111.0.0/22 dev ens192 proto kernel scope link src 10.111.0.32 metric 100
```




Пробуем решить проблему Incast своими силами

Меняем RTO(min)

```
ip route add 10.100.0.0/24 via 10.111.3.254 dev ens192 rto_min 10ms

[root@host test]# ip r

default via 10.111.3.254 dev ens192 proto static metric 100
10.0.0.0/8 via 10.111.3.254 dev ens192 proto bird metric 32
10.111.0.0/22 dev ens192 proto kernel scope link src 10.111.0.32 metric 100

10.100.0.0/24 via 10.111.3.254 dev ens192 rto_min lock 10ms
```




Пробуем решить проблему Incast своими силами

Динамическая маршрутизация и RTO(min)

```
[root@host test]# ip r

default via 10.111.3.254 dev ens192 proto static metric 100
10.0.0.0/8 via 10.111.3.254 dev ens192 proto bird metric 32 rto_min lock 10ms
10.25.0.0/8 via 10.111.3.254 dev ens192 proto bird metric 32 rto_min lock 7ms
10.42.0.0/8 via 10.111.3.254 dev ens192 proto bird metric 32 rto_min lock 5ms

10.111.0.0/22 dev ens192 proto kernel scope link src 10.111.0.32 metric 100 rto_min lock 3ms
```

RTO(min) можно настроить, например, с помощью демона динамической маршрутизации bird
https://bird.network.cz/?get_doc&v=16&f=bird-6.html



Пробуем решить проблему Incast своими силами

Динамическая маршрутизация, RTO(min) в bird

```
…..
filter 'fltr_network_10_24’ {
    if ( net = 10.0.0.0/24 && dest != RTD_UNREACHABLE ) then { krt_lock_rto_min = true; krt_rto_min
=  10; accept; }
}
…..
protocol kernel {
   ipv4 {
      …..
      export filter 'fltr_network_10_24';
   };
}
…..
```




Пробуем решить проблему Incast своими силами

Динамическая маршрутизация, RTO(min) в bird
BGP community

```
function get_rto_from_community() {
  if (65530, 2) ~ bgp_community then { return 2; }
  else return 10;
}
…..
filter 'fltr_network_10_24' {
     if ( net = 10.0.0.0/24 && dest != RTD_UNREACHABLE && defined(bgp_community) && bgp_community
~ [(65530,*)] ) then { krt_lock_rto_min = true; krt_rto_min = 10; accept; }
}
…..
protocol kernel {
    ipv4 {
        …..
        export filter 'fltr_network_10_24';
    };
}
```

Пробуем решить проблему Incast своими силами

Какие проблемы мы можем получить при изменении RTO(min)?

Sender
Receiver
RTO
Retransmission

RTO мало
Ненужный повтор передачи

Sender
Receiver
RTO

RTO большое
Медленная реакция на потери

Подробнее тут https://www.pdl.cmu.edu/PDL-FTP/Storage/sigcomm147-vasudevan.pdf



Пробуем решить проблему Incast своими силами

Что дальше?



Благодарю за внимание



Красников Валерий Викторович

ПАО Сбербанк, SberData
Руководитель направления
Поддержка сервисов кибербезопасности

